Compositional Generative Networks: Approximate Analysis by Synthesis

Alan Yuille

Based on work in 2020 and 2021

Analysis by Synthesis

- Formulates vision in a Bayesian perspective.
- Generative Model P(I|W) – model for generating an image I from a state of the world W
- Prior Model P(W) –state of the world
- Posterior Model P(W|I).
- Synthesis means that we have models P(I|W) P(W) for generating images.
- Analysis means that we can determine the world state W from P(W|I).

Testing Vision Models

- Usually vision models are tested by assuming that we have i.i.d. samples from some unknown distribution.
- We train on half the data and test on the other half. This yields standard performance measures.
- Claim: these are problematic for many reasons.
- Instead – what tougher tests like out-of-distribution, domain transfer, and others.

Compositional Generative Networks

- Generalization under Partial Occlusion

- A Deep Architecture with Innate Robustness to Partial Occlusion
  - A Generative Compositional Model of Neural Features
  - Robustness to Occlusion and Occluder Localization

- Robust Object Detection under Occlusion with CompositionalNets
  - Disentanglement of Context and Object Representation

- Conclusion

Motivation – Generalization under occlusion is important

- In natural images objects are surrounded and partially occluded by other objects
- Occluders are highly variable in terms of shape and texture -> **exponential complexity**
- Vision systems must generalize in exponentially complex domains

Motivation – A Fundamental Limitation of Deep Nets

- DCNNs do not generalize when trained with non-occluded data




Occ. Area	0%	30%	50%	70%	Avg
VGG -16	99.1	88.7	78.8	63.0	82.4

- What if we train with lots of augmented data?



Occ. Area	0%	30%	50%	70%	Avg
VGG-16-Augmented	99.3	92.3	89.9	80.8	90.6

Overview

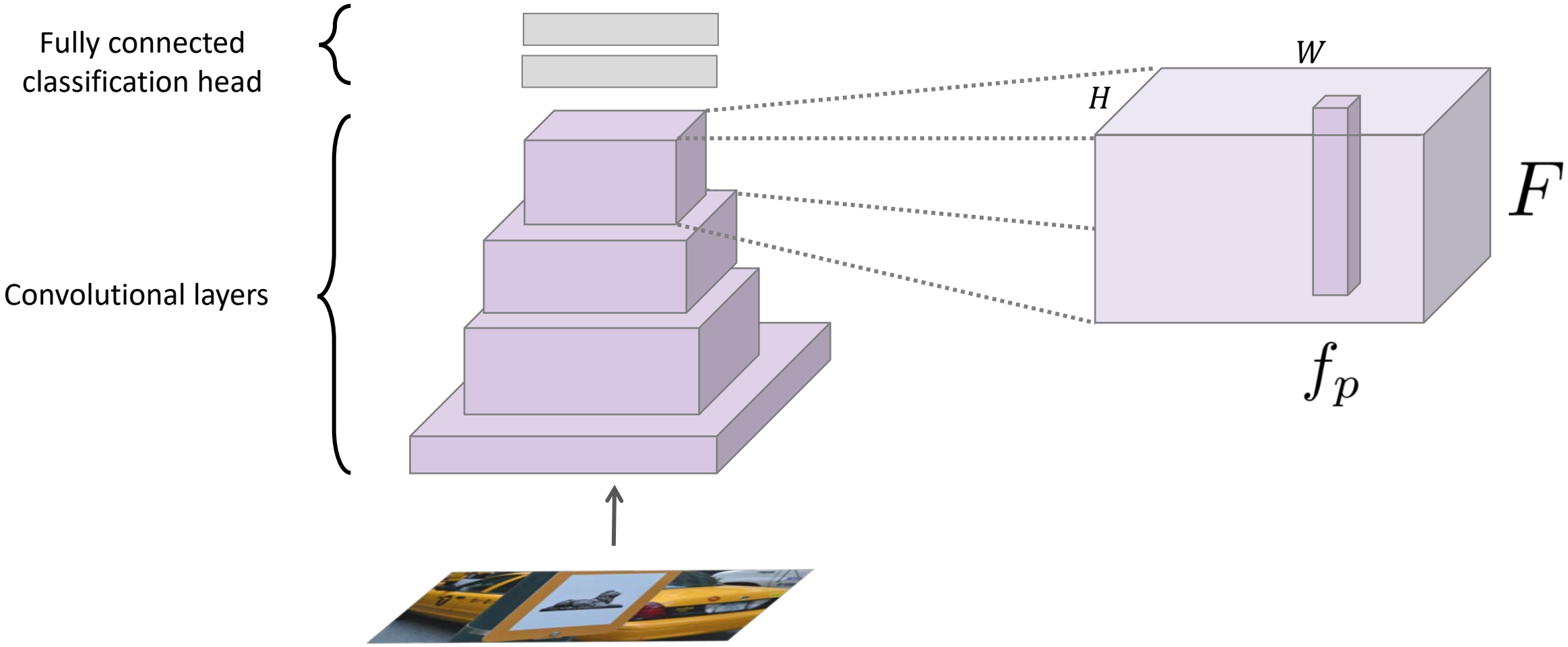
- Generalization under Partial Occlusion

- **A Deep Architecture with Innate Robustness to Partial Occlusion**
  - Generative Compositional Model of Neural Features
  - Robustness to occlusion and occluder localization

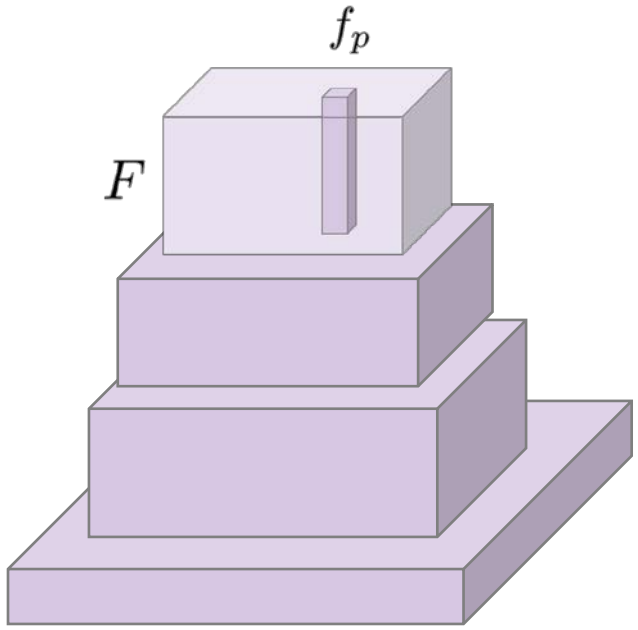
- Robust Object Detection under Occlusion with CompositionalNets
  - Disentanglement of Context and Object Representation

- Conclusion

A Generative Model of Neural Feature Activations



A Generative Model of Neural Feature Activations

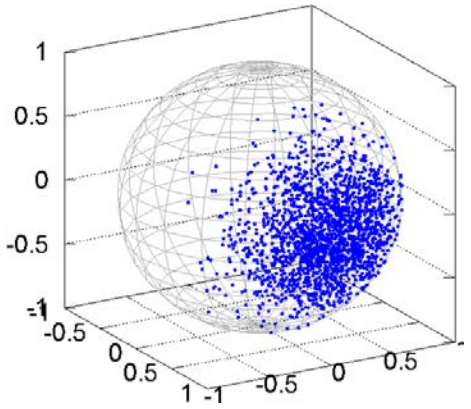


$$p(F|\Theta_y) = \sum_m \nu^m p(F|\theta_y^m)$$

$$p(F|\theta_y^m) = \prod_p p(f_p|\mathcal{A}_{p,y}^m, \Lambda)$$

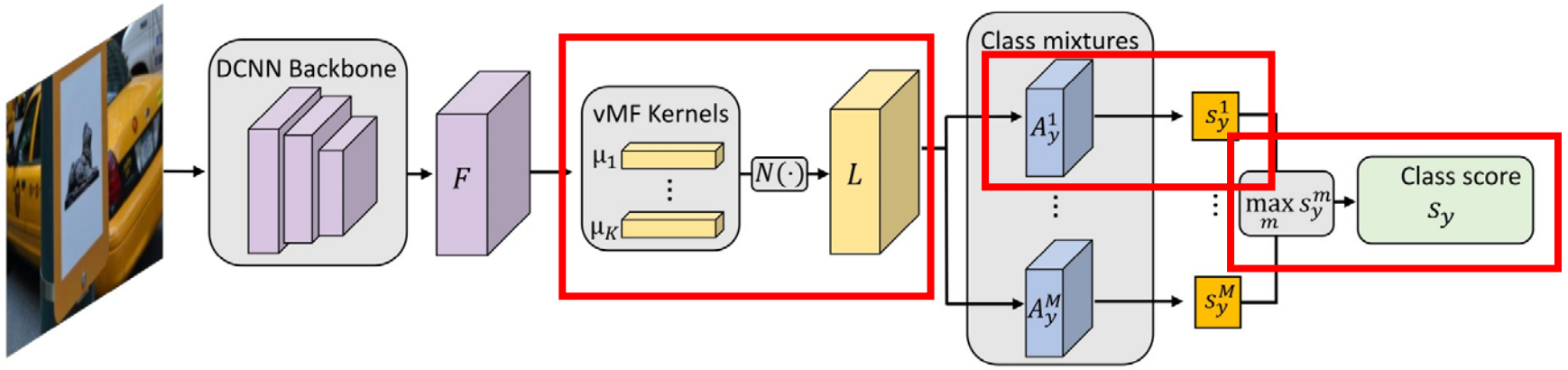
$$p(f_p|\mathcal{A}_{p,y}^m, \Lambda) = \sum_k \alpha_{p,k,y}^m p(f_p|\lambda_k), \quad \lambda_k = \{\mu_k, \sigma_k\}$$

$$p(f_p|\lambda_k) = \frac{e^{\sigma_k \mu_k^T f_p}}{Z(\sigma_k)}, \|f_p\| = 1, \|\mu_k\| = 1$$



Y labels object class
P labels position in the image
fp are the feature vectors at p
m label the mixture (viewpoint)
alpha's,lambda's,mu's are parameters which are learnt.

Inference as Feed-Forward Neural Network

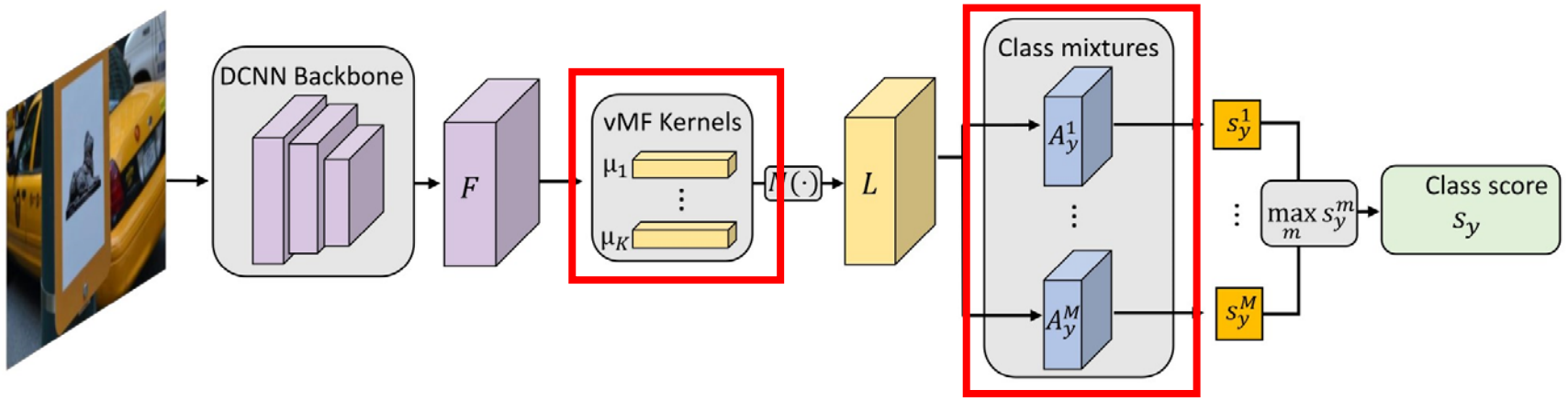


1. vMF likelihood: $$p(f_p|\lambda_k) = \frac{e^{\sigma_k \mu_k^T f_p}}{Z(\sigma_k)}, \|f_p\| = 1, \|\mu_k\| = 1$$

2. Mixture likelihoods: $$p(F|\theta_y^m) = \prod_p \sum_k \alpha_{p,k,y}^m p(f_p|\lambda_k)$$

3. Class score: $$p(F|\Theta_y) = \sum_m \nu^m p(F|\theta_y^m), \quad \nu^m \in \{0,1\}, \sum_m \nu^m = 1$$

Learning the Model Parameters with Backpropagation

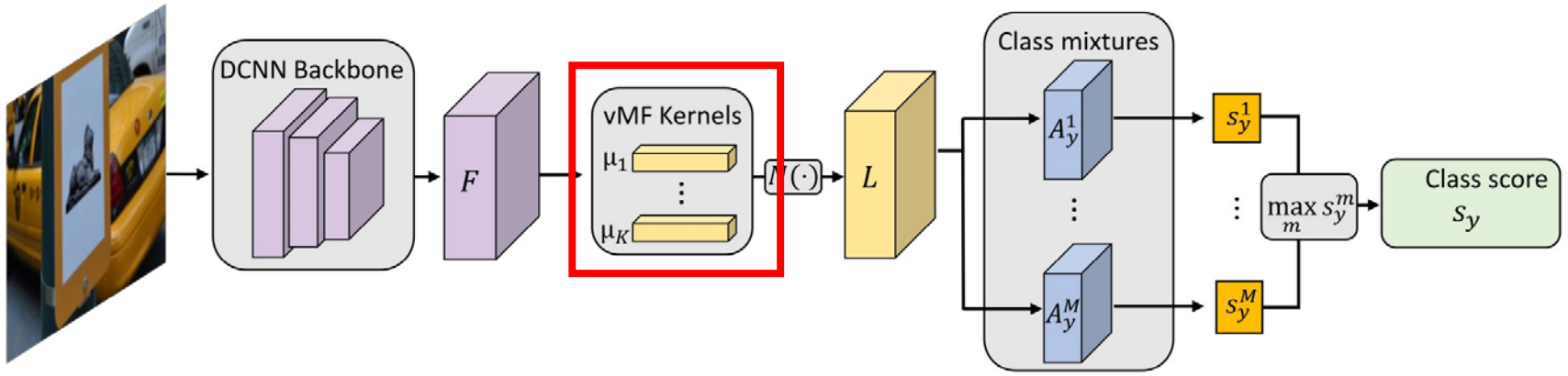


$$\mathcal{L} = \mathcal{L}_{class}(y, y') + \gamma_1 \mathcal{L}_{weight}(W) + \gamma_2 \mathcal{L}_{vmf}(F, \Lambda) + \gamma_3 \mathcal{L}_{mix}(F, \mathcal{A}_y)$$

$$\mathcal{L}_{vmf}(F, \Lambda) = -\sum_p \max_k \log p(f_p|\mu_k) = C \sum_p \min_k \mu_k^T f_p$$

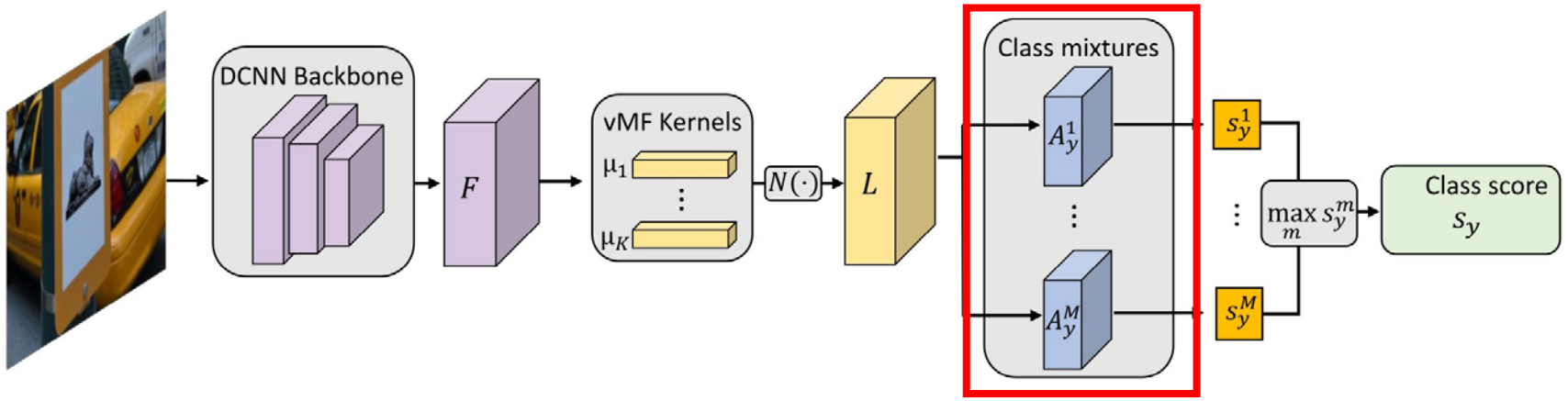
$$\mathcal{L}_{mix}(F, \mathcal{A}_y) = -\sum_p \log \left[\sum_k \alpha_{p,k,y}^m p(f_p|\mu_k) \right]$$

Explainability - vMF Kernels resemble „part detectors“



- Image patterns with highest likelihood:

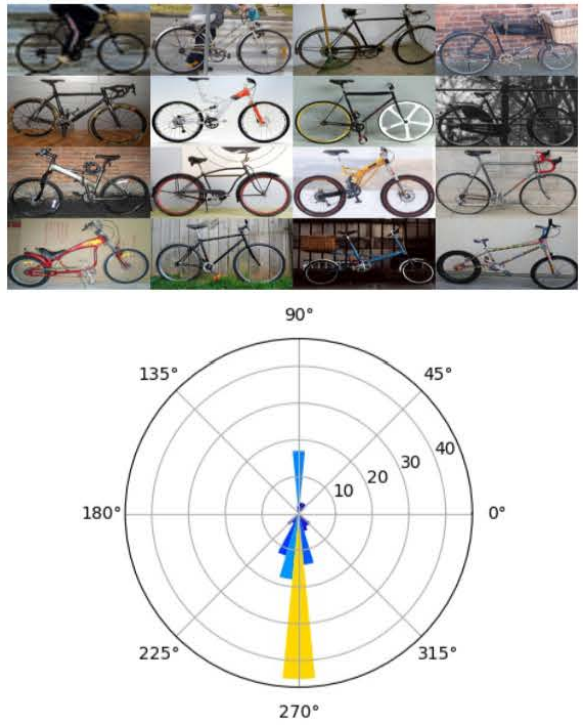
Explainability – Mixture components model object pose

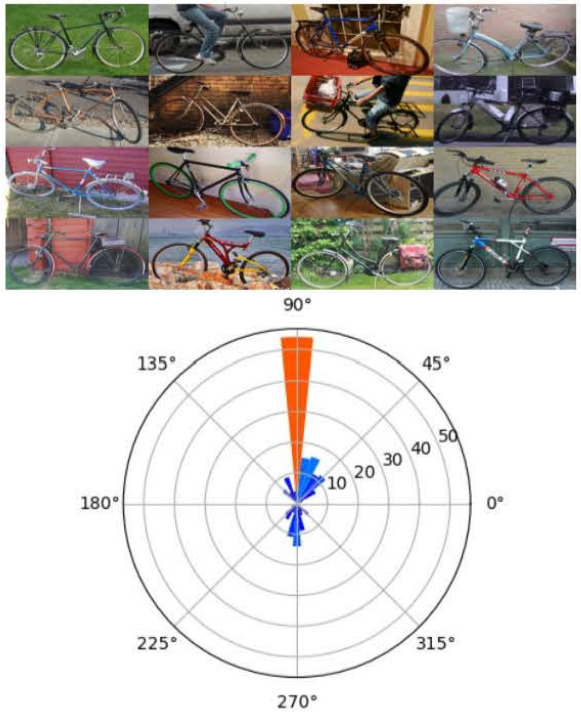


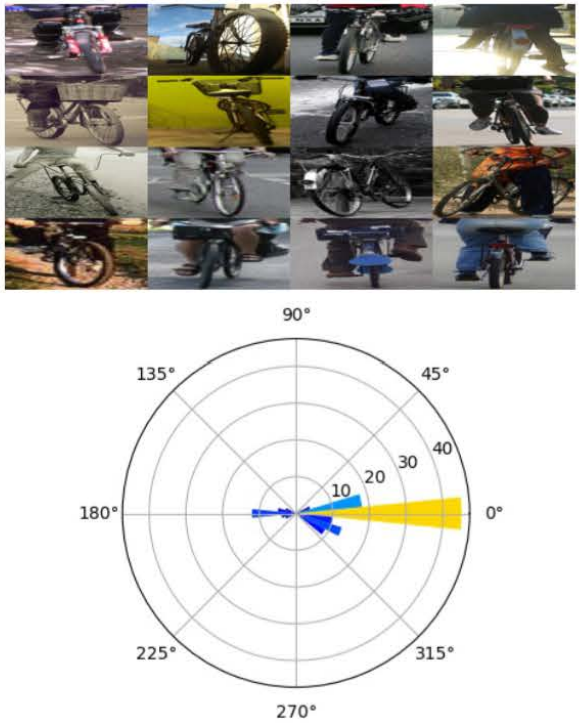
- Images with highest likelihood for mixture components:

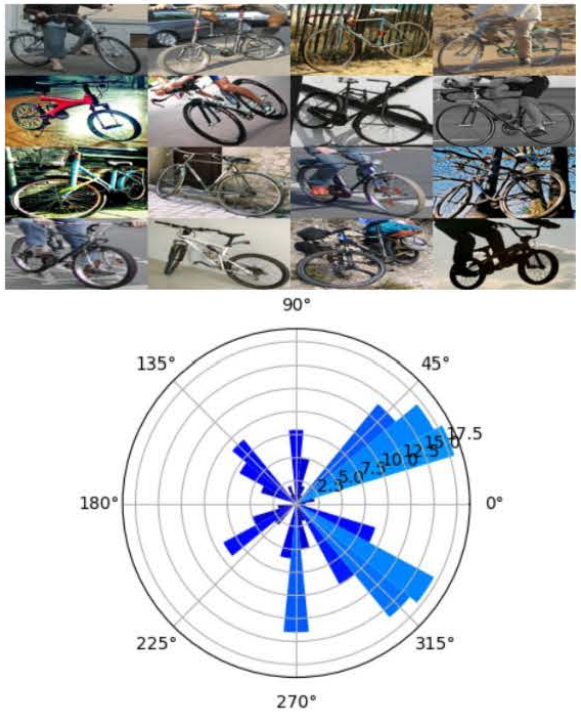



Explainability – Mixture components model object pose

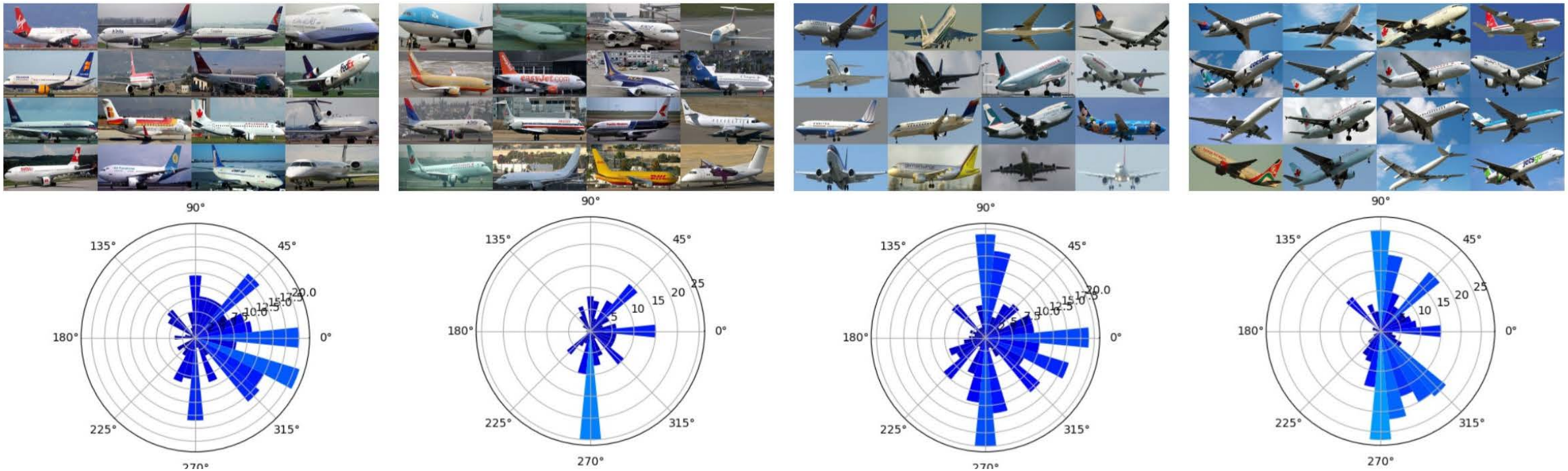








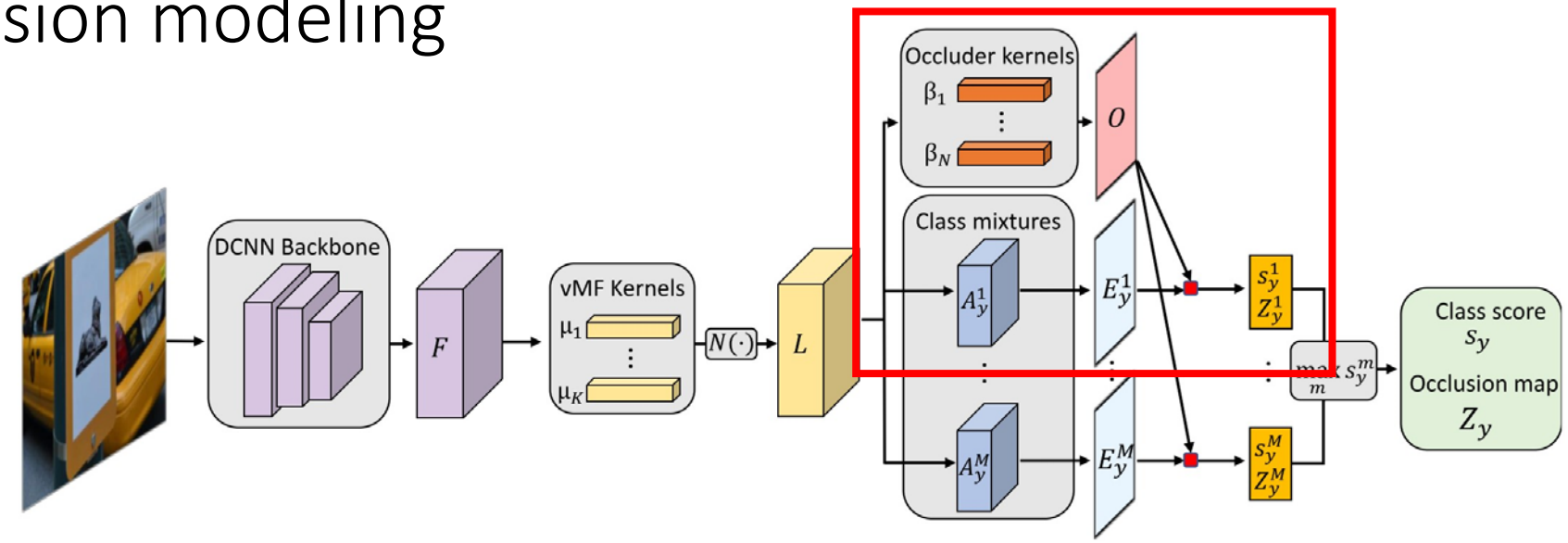
Explainability – Mixture components model object pose



Overview

- Generalization under Partial Occlusion
- A Deep Architecture with Innate Robustness to Partial Occlusion
  - Generative Compositional Model of Neural Features
  - **Robustness to occlusion and occluder localization**
- Robust Object Detection under Occlusion with CompositionalNets
  - Disentanglement of Context and Object Representation
- Conclusion

Occlusion modeling



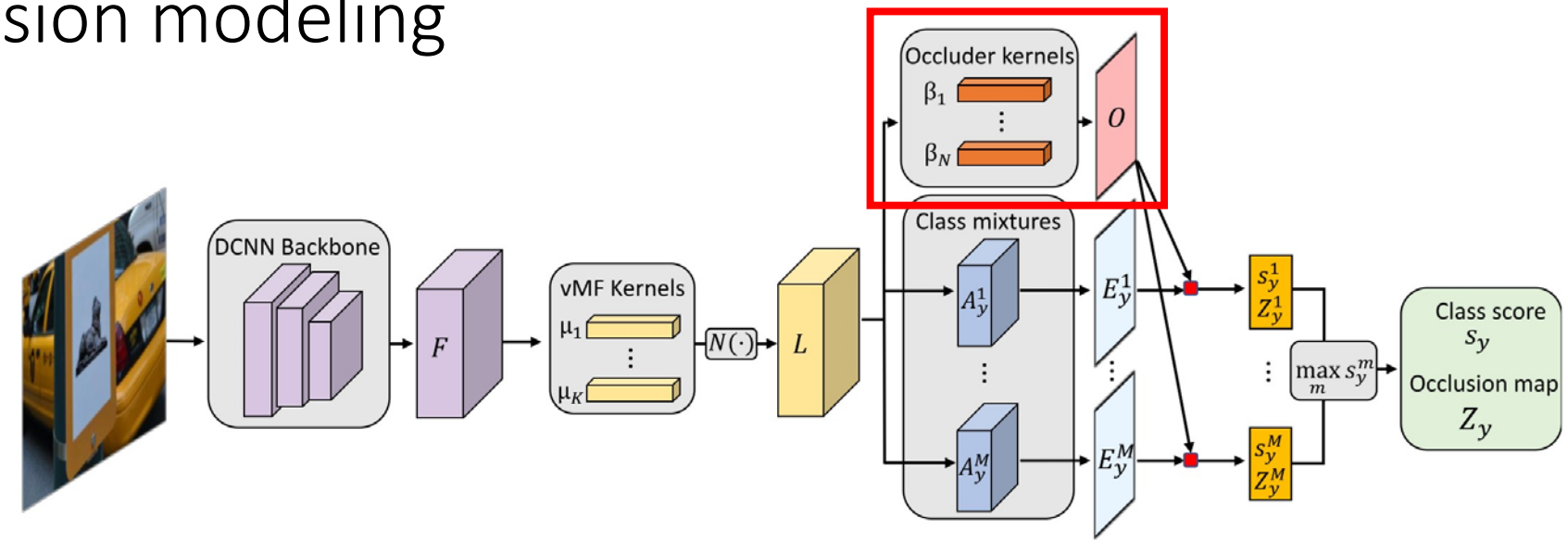
- We introduce an outlier model:

$$p(F|\theta_y^m, \beta) = \prod_p p(f_p, z_p^m = 0)^{1-z_p^m} p(f_p, z_p^m = 1)^{z_p^m}, \quad \mathcal{Z}^m = \{z_p^m \in \{0,1\} | p \in \mathcal{P}\}$$

$$p(f_p, z_p^m = 1) = p(f_p|\beta, \Lambda)\ p(z_p^m = 1),$$

$$p(f_p, z_p^m = 0) = p(f_p|\mathcal{A}_{p,y}^m, \Lambda)\ (1 - p(z_p^m = 1)).$$

Occlusion modeling

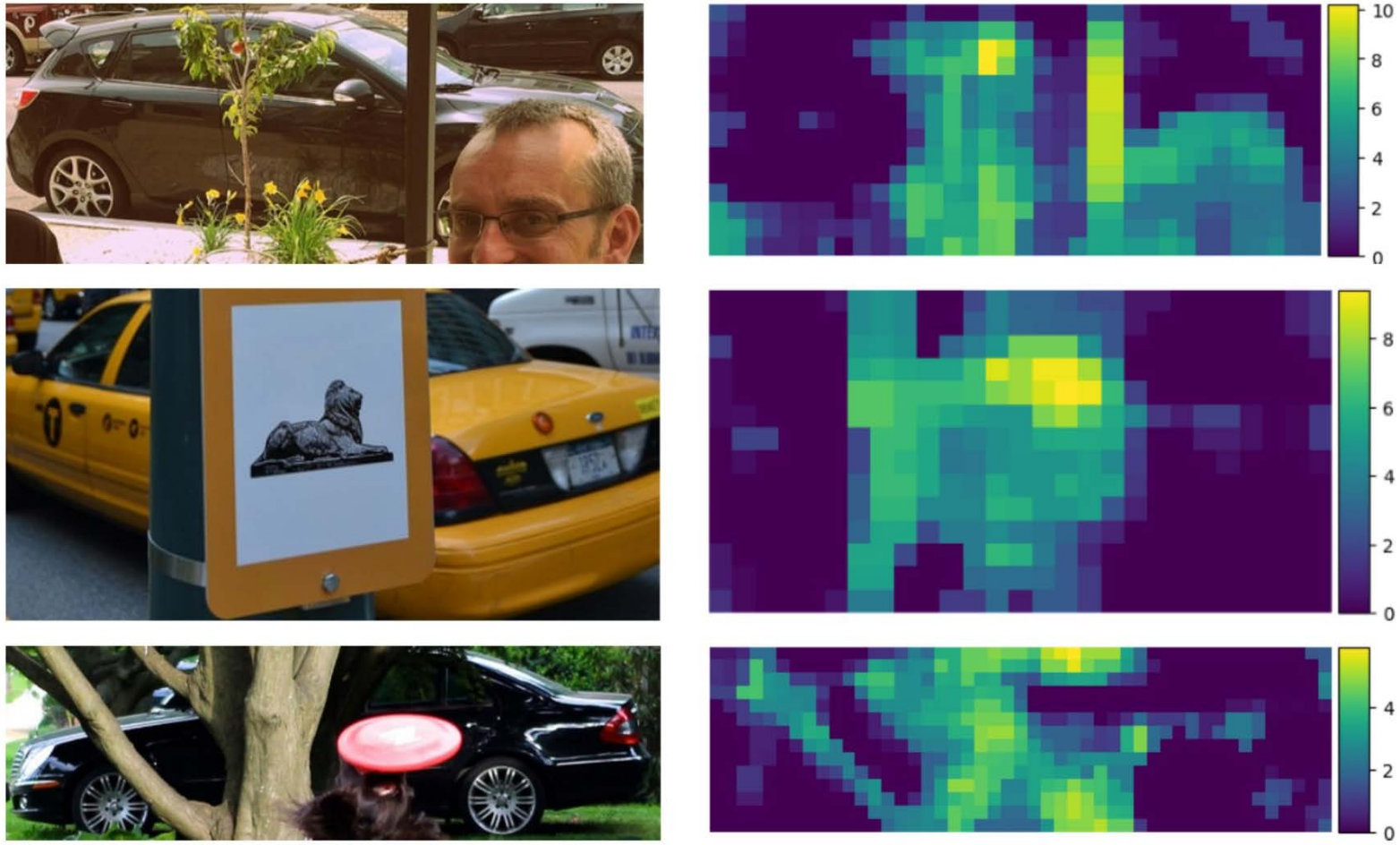


- We introduce an outlier model:

$$p(F|\theta_y^m, \beta) = \prod_p p(f_p, z_p^m = 0)^{1-z_p^m} p(f_p, z_p^m = 1)^{z_p^m}, \quad \mathcal{Z}^m = \{z_p^m \in \{0,1\} | p \in \mathcal{P}\}$$

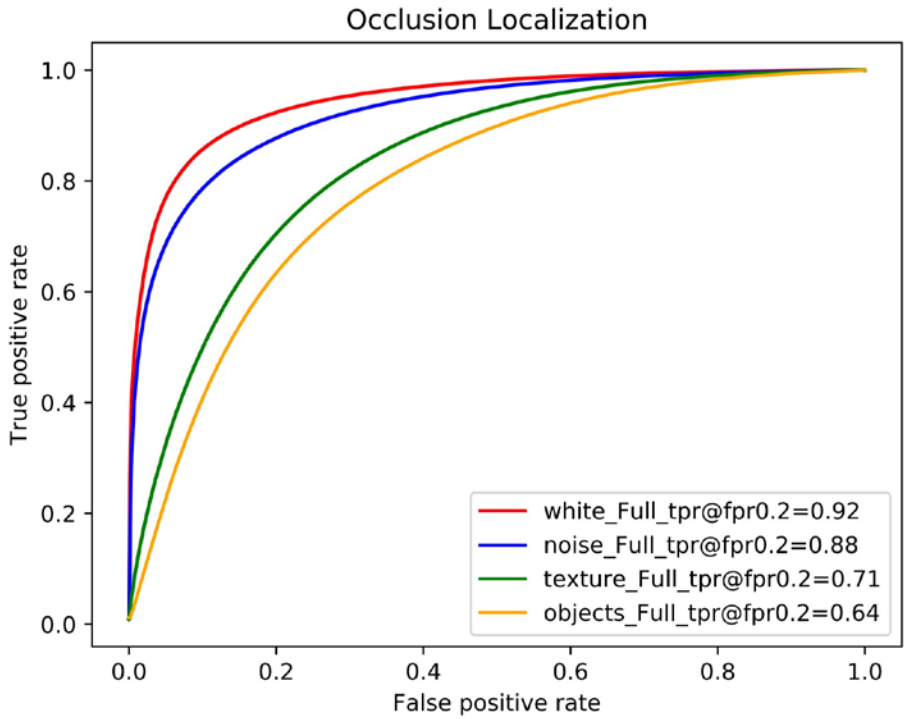
- A simple model of how the object does not look like:

Competition between object and outlier model



Quantitative Evaluation of Occluder Localization



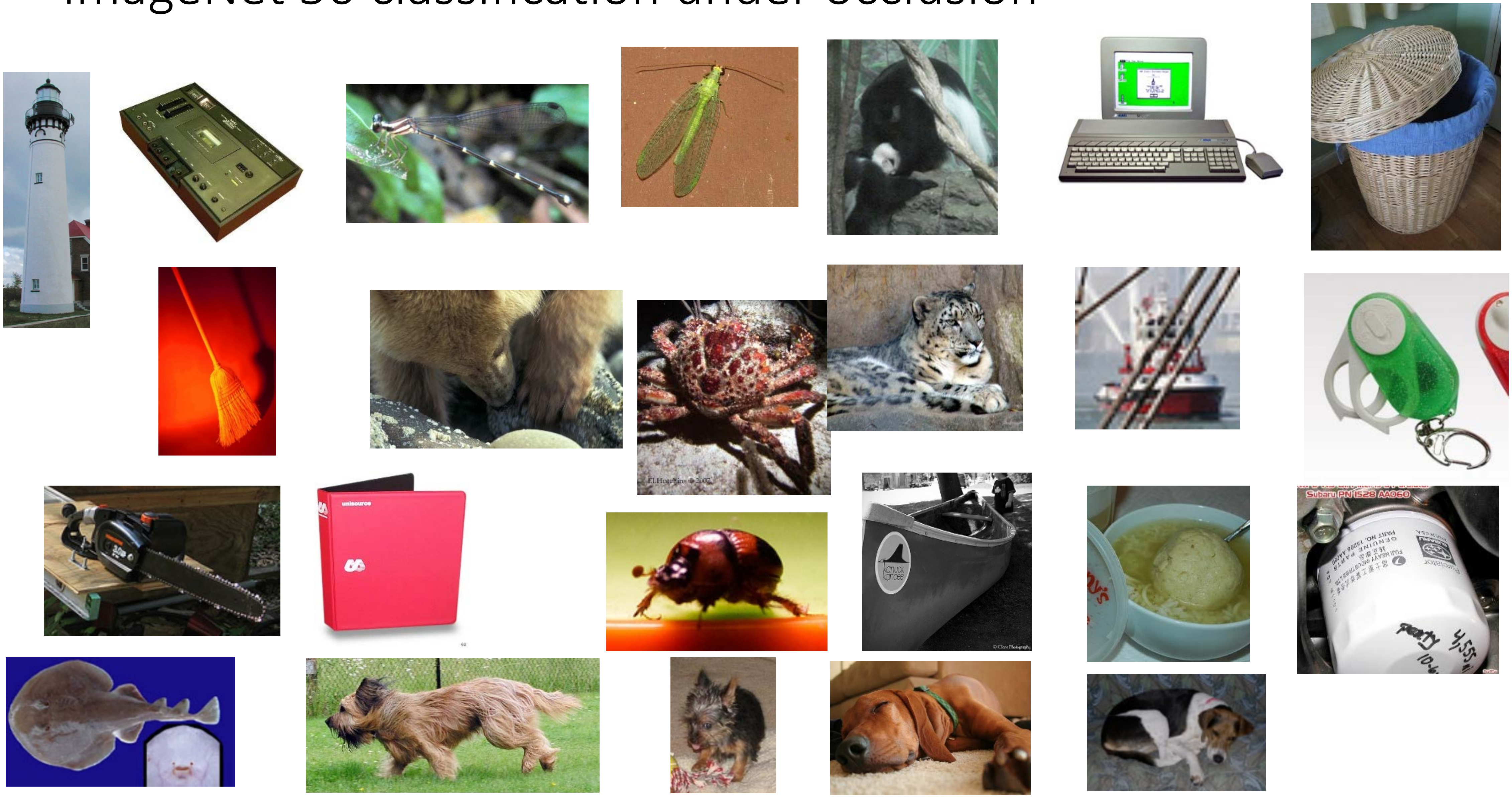



CompNets can classify partially occluded vehicles robustly

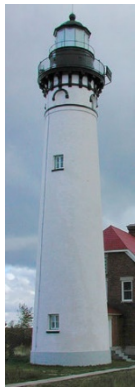


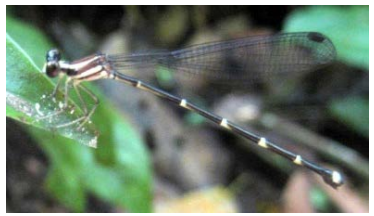

Occ. Area	**L0**	**L1**	**L2**	**L3**	Avg
VGG	97.8	86.8	79.1	60.3	81.0
ResNet50	98.5	89.6	84.9	71.2	86.1
ResNext	98.7	90.7	85.9	75.3	87.7

ImageNet 50 classification under occlusion



ImageNet 50 classification under occlusion





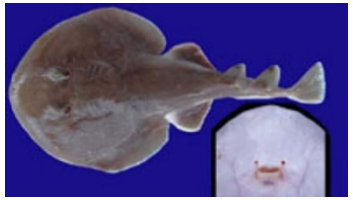










ImageNet under Occlusion

Occ. Area	0%	30%	50%	70%	Avg
ResNext	**98.4**	69.3	48.7	31	61.9
CompNet-ResNext	96.3	**76.6**	**60.1**	**45.5**	**69.6**

Overview

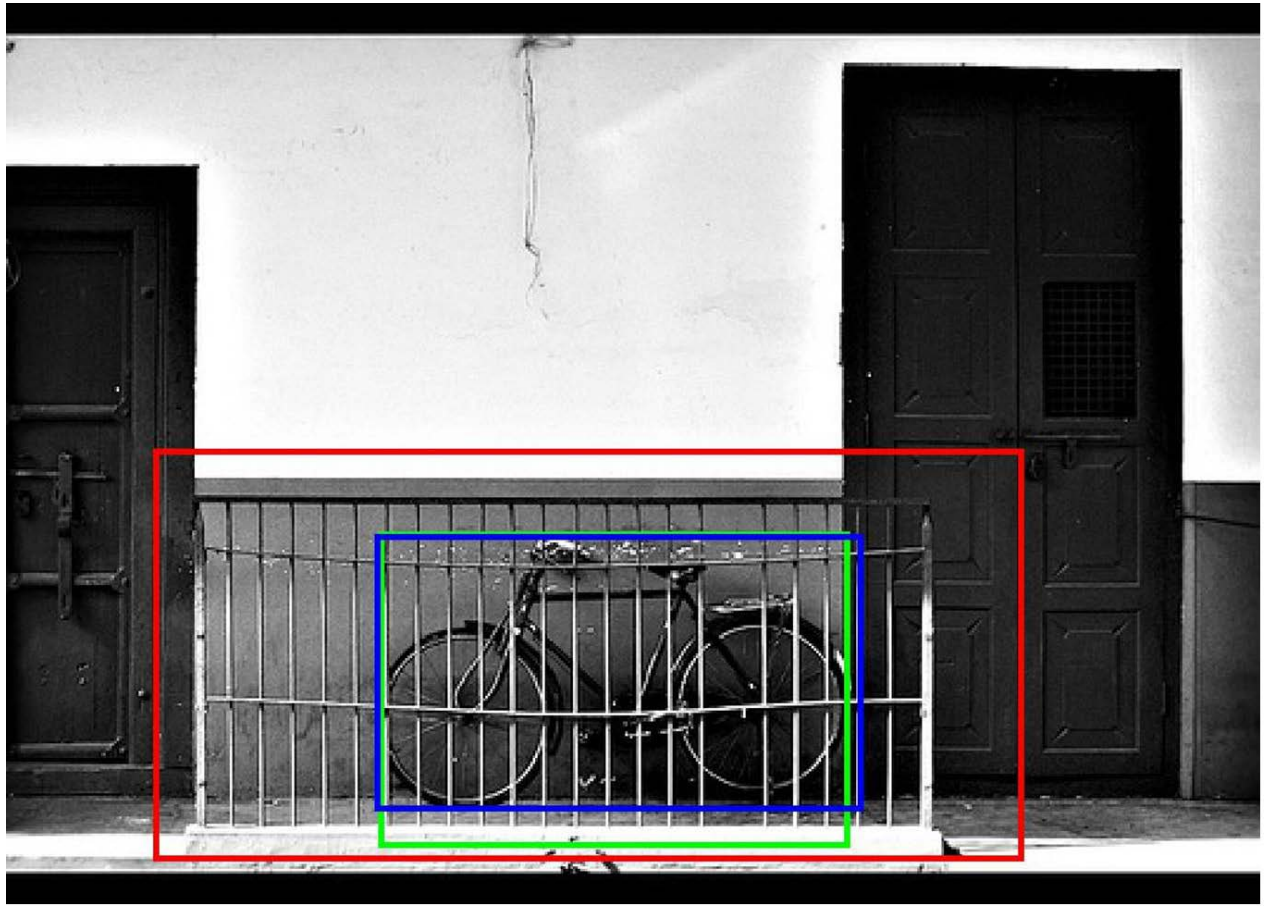
- Generalization under Partial Occlusion

- A Deep Architecture with Innate Robustness to Partial Occlusion
  - Generative Compositional Model of Neural Features
  - Robustness to occlusion and occluder localization

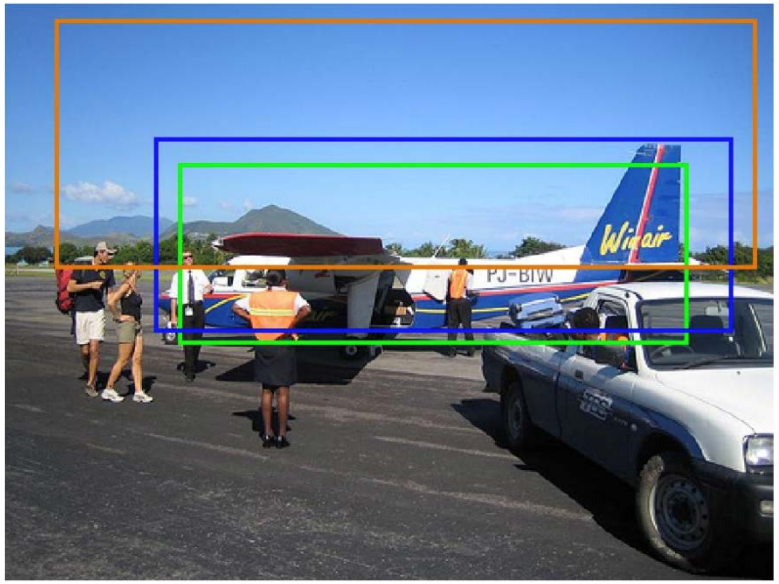
- **Robust Object Detection under Occlusion with CompositionalNets**
  - Disentanglement of Context and Object Representation

- Conclusion

DCNNs for object detection also do not generalize well



Context has too much influence when object is occluded



Seperate the representation of context and object

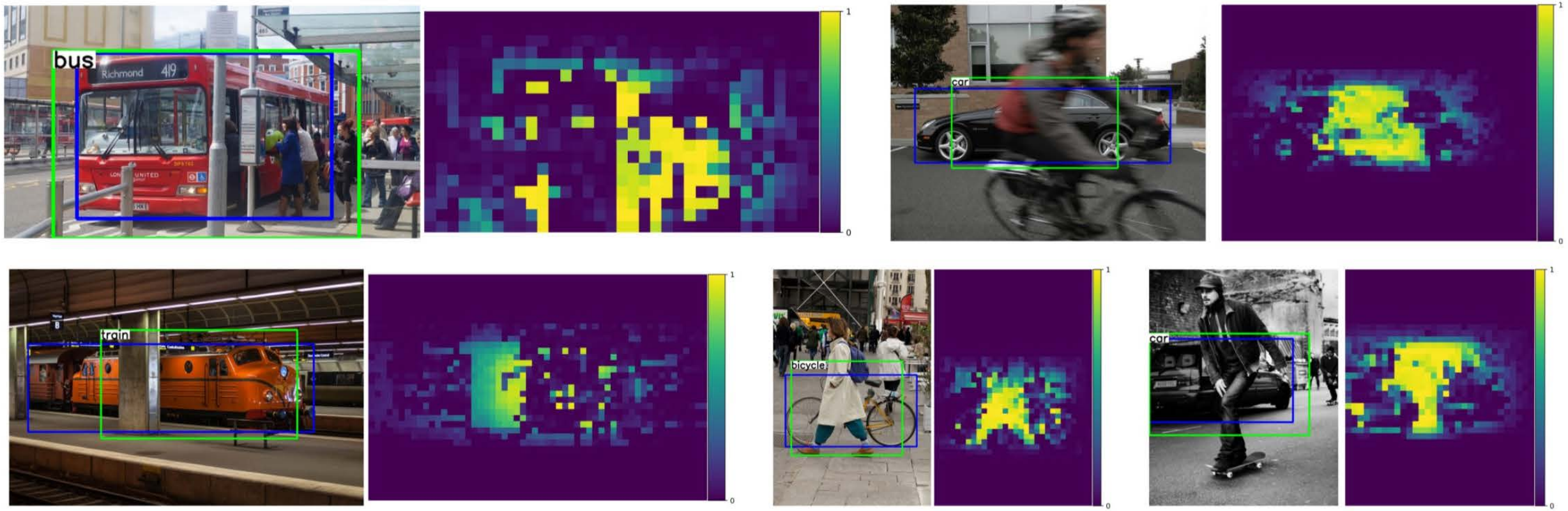
- We introduce a context-aware object model:

$$p(f_p|\mathcal{A}^m_{p,y}, \chi^m_{p,y}, \Lambda) = \omega\ p(f_p|\chi^m_{p,y}, \Lambda) + (1-\omega)p(f_p|\mathcal{A}^m_{p,y}, \Lambda)$$

- Segment the image during training:

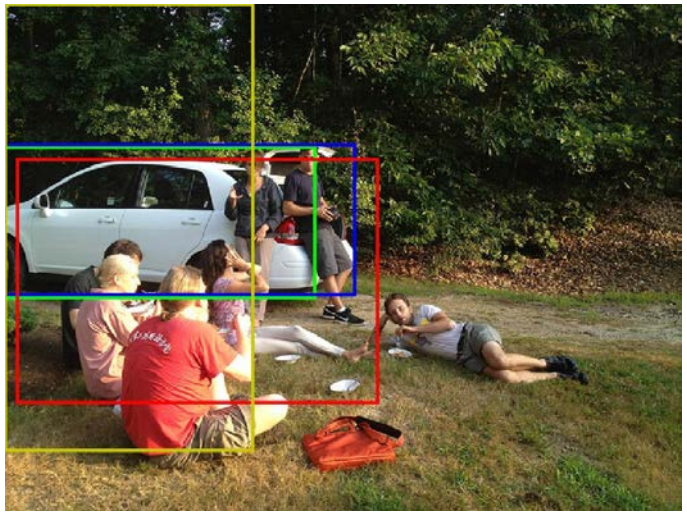
Context-awareness Improves Localization

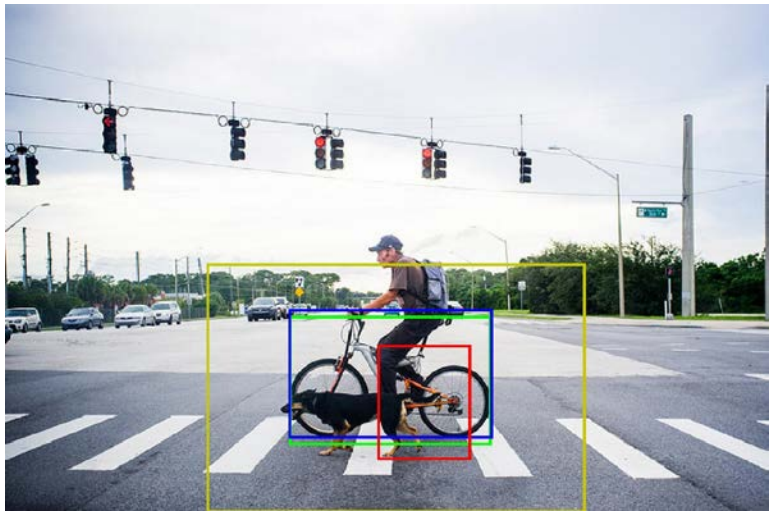
Explainability- Occluder localization in Object detection

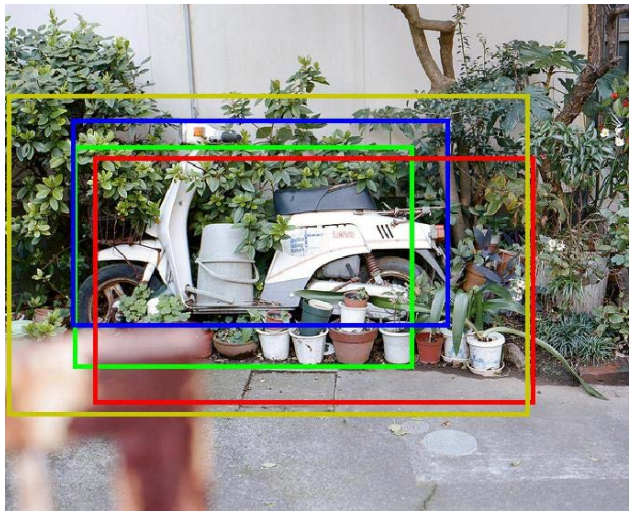


Detection Results

method	light occ.	heavy occ.
Faster R-CNN	73.8	55.2
Faster R-CNN with reg.	74.4	56.3
Faster R-CNN with occ.	77.6	62.4
CA-CompNet via BBV $\omega = 0.5$	78.6	76.2
CA-CompNet via BBV$\omega = 0.2$	**87.9**	**78.2**
CA-CompNet via BBV $\omega = 0$	85.6	75.9







Conclusion

- Partial occlusion introduces exponential complexity in the data
- The complexity gap can be overcome by introducing prior knowledge about compositionality, partial occlusion and context into the neural architecture
- Generalization beyond the training data in terms of partial occlusion & context
- Retain high discriminative performance due to end-to-end training
- Future work: Articulated objects, 3D geometry, top-down reasoning, scale, …

Four Additional Projects

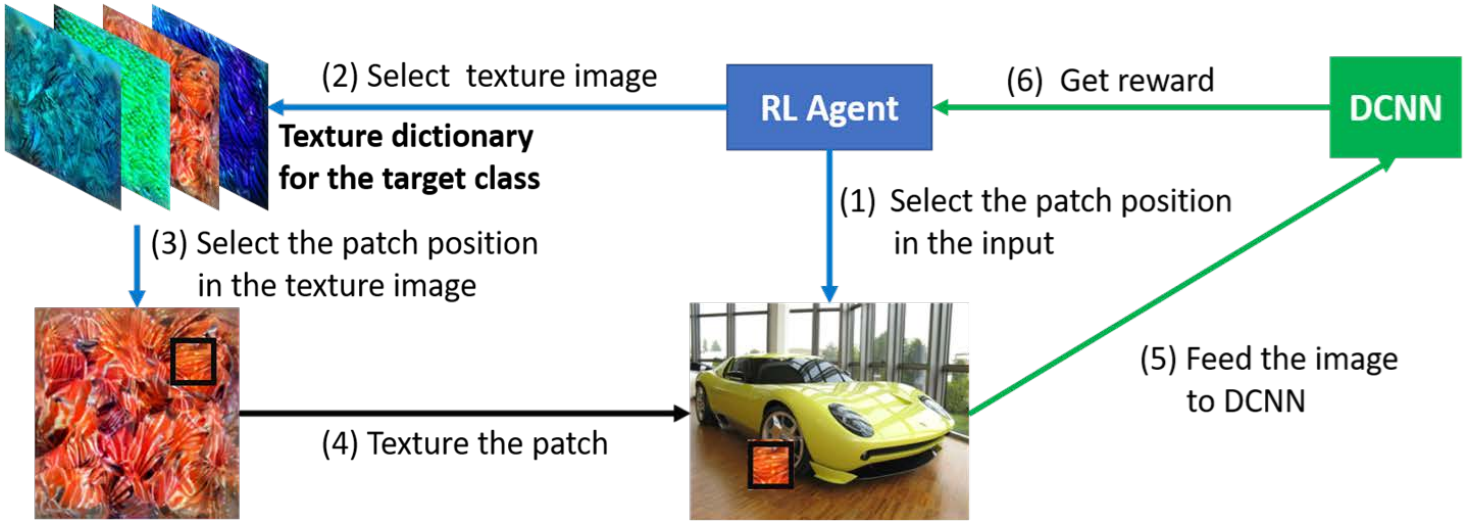
- *1. Robustness to Patch-Attacks. Deep Nets can fail badly when stress-tested by Patch-Attacks, but CompNets are robust out of the box.* (Christian Cosgrove at al. In submission. 2020).
- *2. Multi-task consistency – modal and amodal boundary detection without boundary annotations.* (Yihong Sun et al. In submission, 2020).
- *3. Recurrent Reasoning about Multi-Object Occlusion. Bottom-up and to-down.* (Xiaoding Yuan et al. In submission. 2020).
- *4. NeMo – neural mesh model for robust 3D pose estimation.* (Angtian Wang et al. In submission. 2020).

(1): ROB. Robustness out of the box

- Background – standard computer vision & machine learing practice is to evaluate algorithms by average case performance on a finite-sized balanced annotated dataset (BAD).
- We argue that it is better to evaluate algorithms by trying to identify their weak points by dynamic testing, i.e. modifying the input images adaptively to cause the algorithms to fail – Adversarial Examiner.
- In previous work, Chenglin Yang et al. CVPR. 2020, we developed patch-attacks which could fool Deep Nets by adding a few small patches to images. The patches are their locations were chosen by a search strategy with feedback from the algorithm. *Presented to CompCogSci in Spring 2020.*
- *Dataset: Pascal+PatchAttacks. German-Traffic-Signs+PatchAttacks.*
- These blackbox targeted attacks had over 90% success rate on advanced Deep Nets. Suggests that Deep Nets lack knowledge of the spatial structures of objects.

(1) ROB Patch Attacks: Overview

- Learn an Attack Policy by reinforcement learning.



Network	Attack	T_acc. (%)	Avg_area (%)	Avg_qry
ResNet50	–	0.10	–	–
	HPA	23.20	71.54	50000
	MPA_RGB	25.90	18.45	28361
	TPA_N10_2%	97.60	7.80	15728
	TPA_N10_4%	99.70	9.97	8643
	TPA_N10_10%	100.00	15.36	3747

- Chenglin Yang et al. Patch Attack. ECCV. 2020.

(1) ROB: Robustness out of the Box.

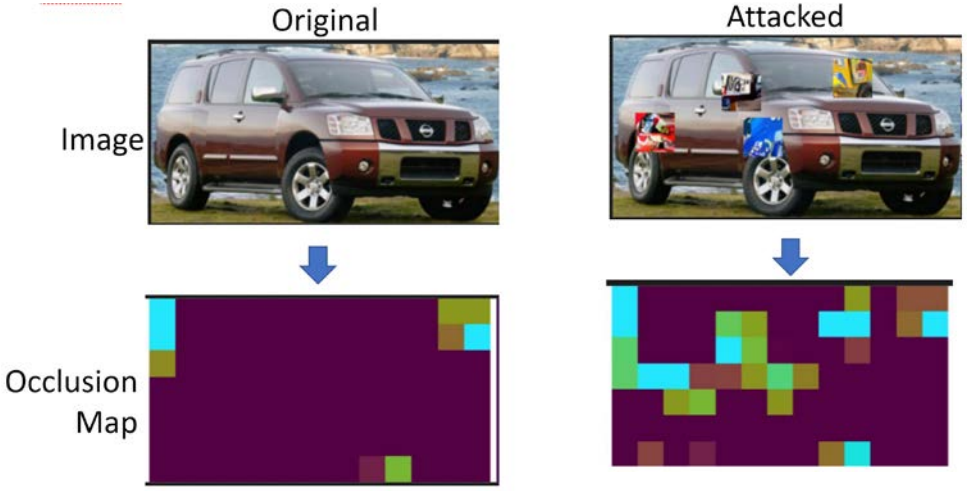
- We conjectured that CompNets would do better than Deep Nets because they have knowledge of the spatial structure of objects and their outlier process may enable them to reject the attacking patches.
- This was correct. Patch-attacks (and related attacks) are less successful on CompNets by an order of magnitude. CompNets also have some ability to detect and localize the patch attacks.
- This gives more evidence that CompNets are much more robust than Deep Nets. Datatset: PASCAL+PatchAttacks.

(1) ROB CompNets are robust to Patch Attacks.

- CompNets ***are*** robust against targeted patch attacks.

Model	Accuracy (%)	Attack success rate (%)	
		PatchAttack[1] (TPA) 4 patches	Sparse-RS[2] 1 patch
CompNet (vgg16 backbone)	98.5	12.6	0.9
vgg16	98.6	98.8	92.3

CompNets can localize attacks.



- C. Cosgrove et al. Robustness out of the box. Arxiv. 2020.

(1) ROB: Need to modify CompNets for fine detail

- We also tested the robustness of CompNets to patch attacks on the *German Traffic Sign dataset*.
- Performance was better than alternative methods, particularly for signs which differed at the coarse level. But CompNets were less effective for fine scale discrimination (e.g., speed limit 60 mph vrs. speed limit 80 mph). We used an engineering trick to partially solve this problem in the paper.
- *Deeper Understanding.* CompNets are generative on features at the upper levels of the Deep Net. These features are invariant to low-level details of the image. This has a big advantage because modeling these details is hard and they are often irrelevant to the task. *But this suggests that we need to supplement our coarse CompNets with fine-scale CompNets defined at lower-levels and with more detailed geometry.*
- This is another twist on the idea of bottom-up processing seeks invariant features while top-down is needed for higher resolution (David Mumford, Tai Sing Lee).

(1) ROB Summary

- CompNets are much more robust than Deep Nets to patch-attacks without needing any modifications.
- Performance degrades for fine-detail tasks (can be fixed by engineering tricks in the short term).
- This requires creating a theory which can model objects at different levels of resolution with only coarse geometry needed at the higher level and more precise and detailed geometry at the lower-levels.
- This also motivates adversarial testing of CompNets, and other models, with adversarial examiners which aim to target their weak points (IAA grant with APL).

(2): Modal and Amodal Boundaries (MAB)

- CompNets have a natural ability to detect object boundaries without any annotated segmentation. They are trained, like Deep Nets, on bounding boxes contain foreground objects and background.
- They learn generative models for the bounding boxes but can also learn generative models for the background. This enables them to make an estimate of the foreground using log-likelihood ratio tests.
- Example from A. Wang et al. 2020.

(2) MAB: Modal and Amodal Boundaries

- The modal boundaries are the ones you can see. The amodal boundaries are invisible.
- Example:

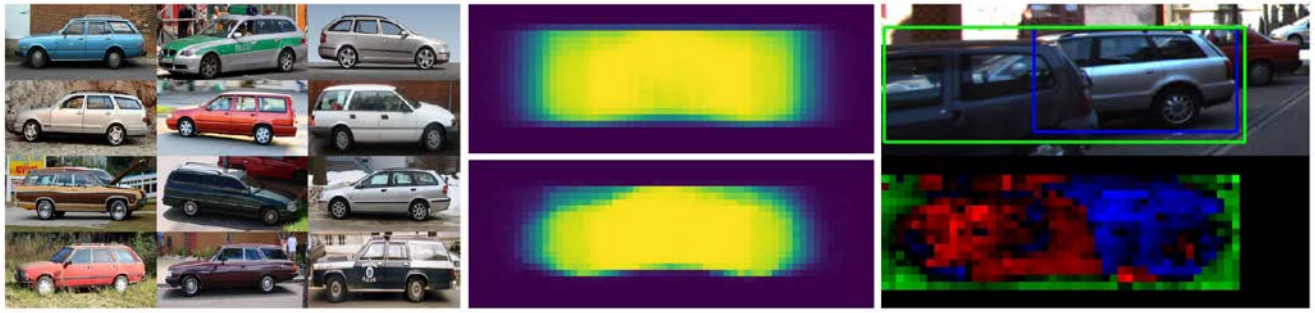
- Note: most people who study Admodal boundary detection assume that the visible segmentations have been annotated (during training). But we do not, so we evaluate CompNets for both Modal and Amodal boundary detection.

(2) MAB: Augment CompNets by a Prior

- We augment CompNets by learning a prior for the boundary of the object, conditioned on the mixture (viewpoint).
- This is done by an EM formulation. We initialize our estimates of the boundaries – both modal and amodal – using the foreground/background estimation as discussed before.
- This prior enables us to predict the amodal boundaries and also improves performance on the modal boundaries.
- We have results on datasets: Kitti-Amodal (we should pay students to annotate Coco-amodal with class labels).

(2) MAB: Results

- Priors and Results. Note: PCNet-M requires supervision (ground truth object segmentation mask). Ours and BBTP does not. Results on KINS dataset (similar results on OccludedVehiclesDetection Dataset).



(a) Object Cluster (b) Compositional Prior (c) Amodal Completion

FG Occ. Level	-	**0**	**1**	**2**	**3**	Mean
PCNet-M	*m*	**89**	**75.1**	49.1	30.5	60.9
ours	*m*	72.1	70.4	**68.4**	**52.7**	**65.9**
BBTP	*a*	73	70.3	66.6	64.7	68.7
ours	*a*	**77.4**	**74.9**	**78.1**	**76.3**	**76.7**

(2) MAB: Summary

- This shows that CompNets can learn the shape of the object despite not being given any annotated segmentation to learn from. This is unlike Deep Nets which simply classifies the annotated bounding boxes.
- This is an example of multi-task-consistency. The CompNet will perform object classification and modal/amodal boundary estimation using the same underlying model (plus some viewpoint/parts).
- This is an example of learning efficiency. The CompNet only requires object annotations (for bounding boxes) but performs well on other visual tasks.
- The distinction between foreground and background can be used for other problems.

(3) M-O-O: Multi-Object Occlusion

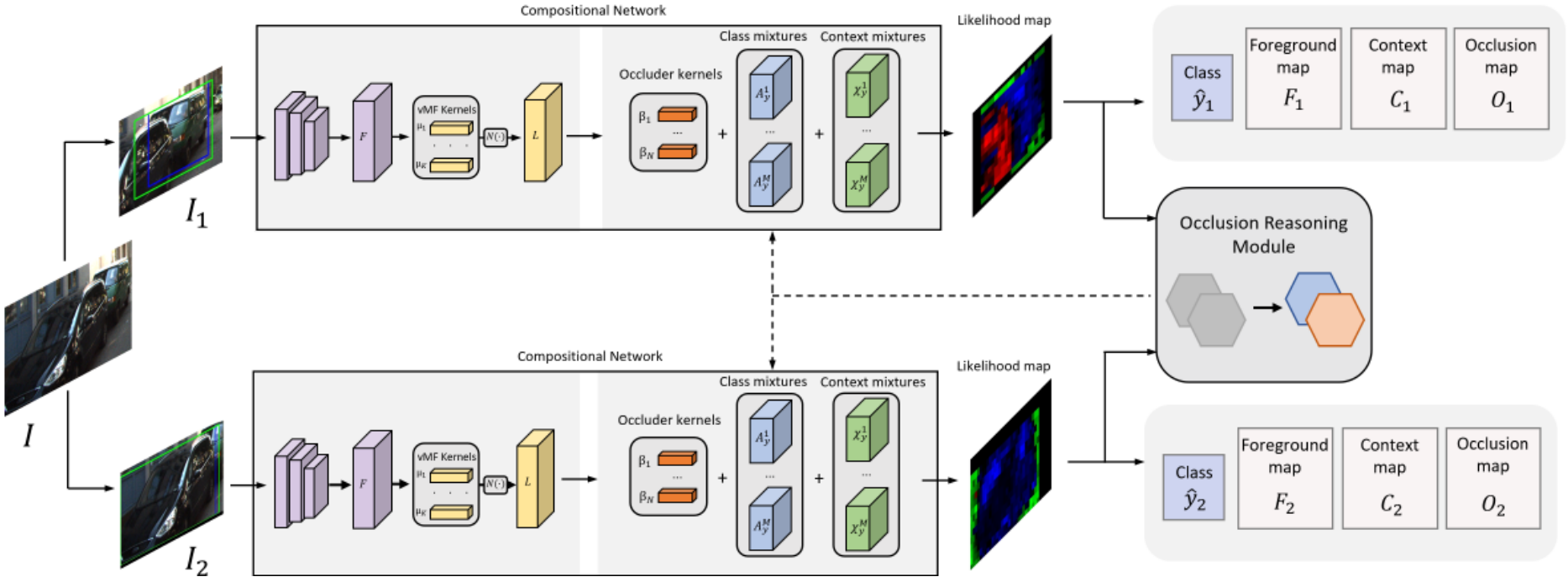
- CompNets assumed there was a single object in the image which could be occluded.
- *We must consider the case where there two, or more, objects, with one object partly occluding the other. We want to detect/classify both objects, segment them, and determine the boundary between them (instance level segmentation).*
- This relates to the Captcha Task.
- *This visual task is hard to do purely bottom-up. We address it by using a bottom-up process (CompNet) to make hypotheses for the partial segmentations of the objects. Then a top-down process, the Occlusion Reasoning Module, finds which interpretations are most consistent with the image.*
- As before, we only use bounding box annotations for the objects.

(3) M-O-O

- We train and evaluate for detecting and segmenting cars on the KITTI dataset. Also a partly synthetic dataset.
- CompNets are augmented with an Occlusion Reasoning Module (ORM). This detects erroneous feed-forward predictions and corrects them by reasoning about the occlusion order of objects.
- We also can handle unknown occluders, which competing methods cannot, and we also extract the occlusion order (i.e. relative depth).

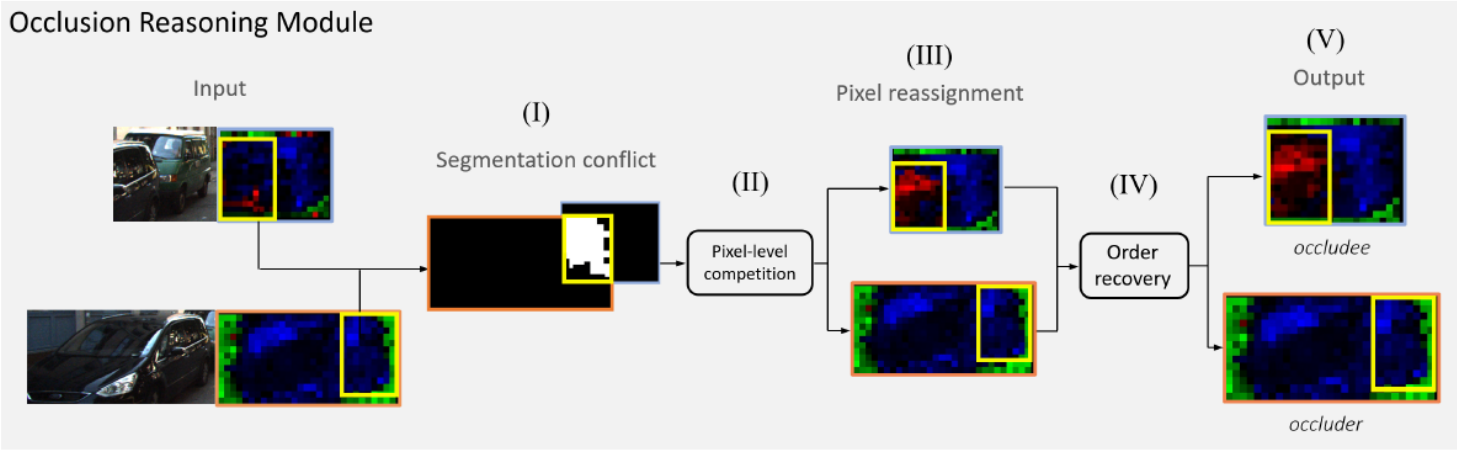
(3) M-O-O Theory

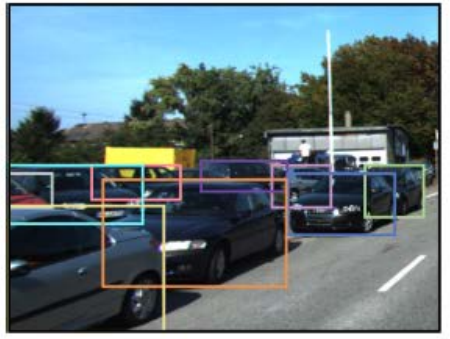
- Our model corrects erroneous instance segmentations by multi-object reasoning.



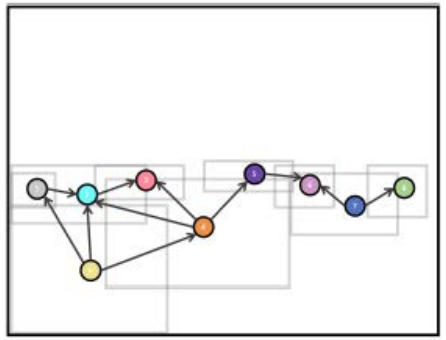
(3) M-O-O More Details

- Detailed model. Resolution of Segmentation conflict.





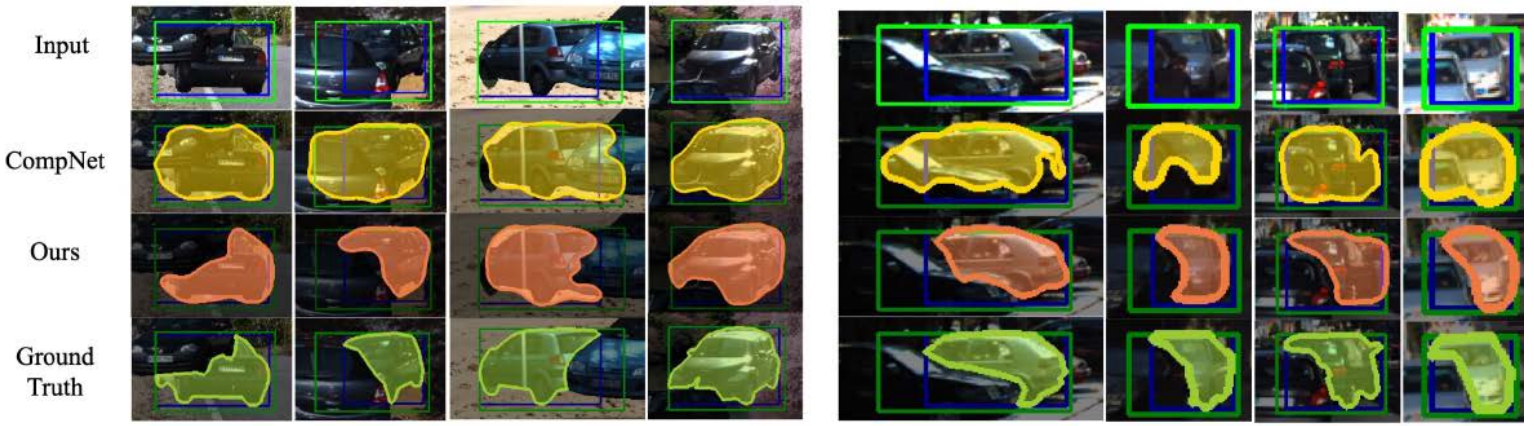
Input image



Order graph

(3) M-O-O Results

- Visualize. And Tables.



(a) Synthetic dataset (b) KINS

	2 Objects					4 Objects					2 Objects + Unknown Occlusion				
Occ Level	L0	L1	L2	L3	Mean	L0	L1	L2	L3	Mean	L0	L1	L2	L3	Mean
Mask R-CNN	88.2	86.3	69.1	58.2	82.3	88.7	88	74.8	63	78.6	90.5	86.8	72.2	57.1	76.7
CompNet	77.8	67.3	51.0	26.3	66.9	**76.7**	67.1	50.2	26.1	56.0	**78.9**	72.2	57.8	36.0	63.6
Ours (iter=1)	**78.0**	**75.3**	65.4	45.6	72.9	75.2	**72.9**	61.9	43.0	65.0	77.9	**73.3**	**62.0**	**41.7**	**65.8**
Ours (iter=2)	**78.0**	**75.3**	**65.7**	**47.2**	**73.1**	75.2	**72.9**	**62.2**	**44.0**	**65.3**	78.0	**73.3**	**62.0**	**41.7**	**65.8**

	2 Objects					4 Objects					2 Objects + Unknown Occlusion				
Occ Level	L0	L1	L2	L3	Mean	L0	L1	L2	L3	Mean	L0	L1	L2	L3	Mean
PCNet-M	82.4	81	69.3	47	70	87.2	79.3	63.7	41.3	67.9	-	-	-	-	-
BBTP	**80.5**	73.6	69.5	72.8	74.1	**80.5**	71.9	64	66	70.6	**83.7**	77.3	67.9	60.6	72.4
CompNet	78.0	76.6	75.0	72.1	76.7	77.3	75.4	74.1	71.4	74.8	78.4	**78.1**	76.1	71.9	76.5
Ours (iter=1)	79.9	**80.0**	79.2	77.7	**79.7**	78.6	78.9	78.1	76.6	78.2	78.6	78.0	**76.2**	**72.1**	**76.6**
Ours (iter=2)	79.9	**80.0**	**79.3**	**78.1**	**79.7**	80.0	**80.0**	**79.3**	**78.1**	**79.5**	78.5	**78.1**	**76.2**	**72.1**	**76.6**

(3) M-O-O Summary

- This project extends CompNets to instance level segmentation on partially overlapping objects.
- We demonstrated this on KITTI with state-of-the-art results.
- This model required bottom-up and top-down processing to implement the Occlusion Reasoning Module. This is simpler than the methods used for CAPTCHA's (George et al. 2017). We will need to extend our model to deal with more challenging conditions.
- This method can be adapted to other types of objects (need to annotate Coco with labels first). Maybe it could also help defend against advanced patch attacks one object occluded by fractions of another object (which obey the correct spatial relationships).

(4) NeMo: Robust 3D Pose Estimation

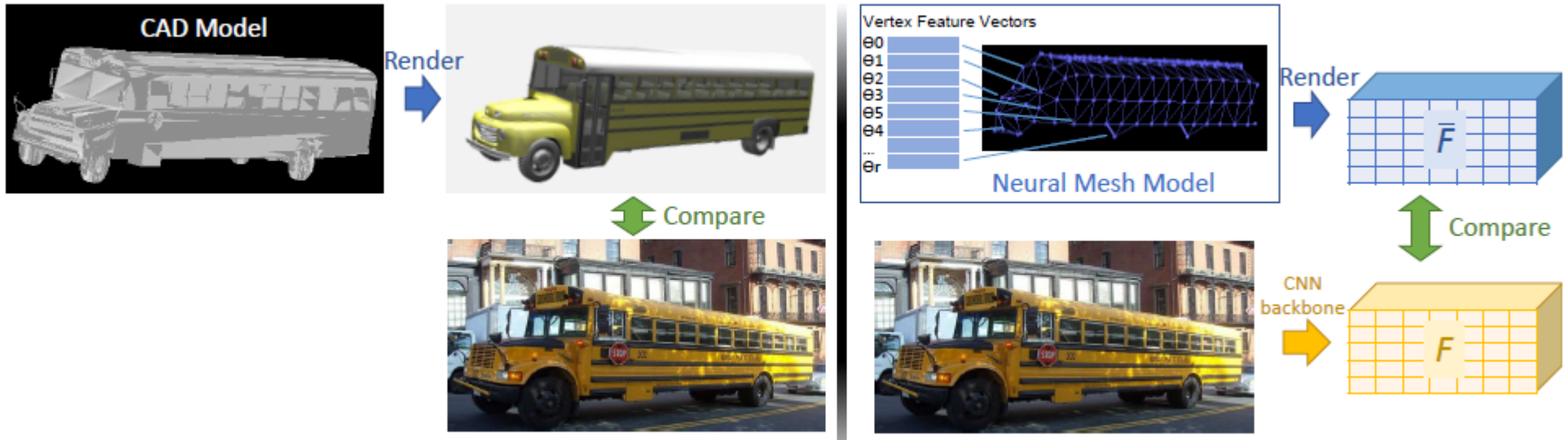
- This is our first attempt to extend CompNets to have 3D representations.
- We use a Mesh model to represent the object. This does not give precise geometry. But it is adequate for our purpose, as we will discuss, and can be refined later.
- We target 3D pose estimation under occlusion because there are datasets and challenges for estimating 3D pose of objects. But nobody has studied robustness to occlusion, which we introduce.
- We can also generalize to novel views (without occlusion).
- Datasets: Pascal3D+, ObjectNet3D.
- Learning Supervision: the 3D mesh models are paired with the images.

(4) NeMo

- This 3D CompNet is generative on either: (i) Deep Network feature vectors, or (ii) feature vectors learnt by contrastive learning.
- *This differs from being generative on the image appearance. We only have to learn the distribution of feature vectors which have been trained so that they are invariant to unimportant details of the object which includes its precise texture appearance and also to its exact shape, which is why a mesh model is adequate (five meshes for different types of cars).*
- This also makes inference easier to search for the 3D viewpoint. We are searching over a coarse representation of the object which is much easier (and we can quantify).
- *This has the practical advantage that we do not need to learn a high-precision generative model for each object, which would be very difficult.*
- Note: Contrastive learning features are better for this task.

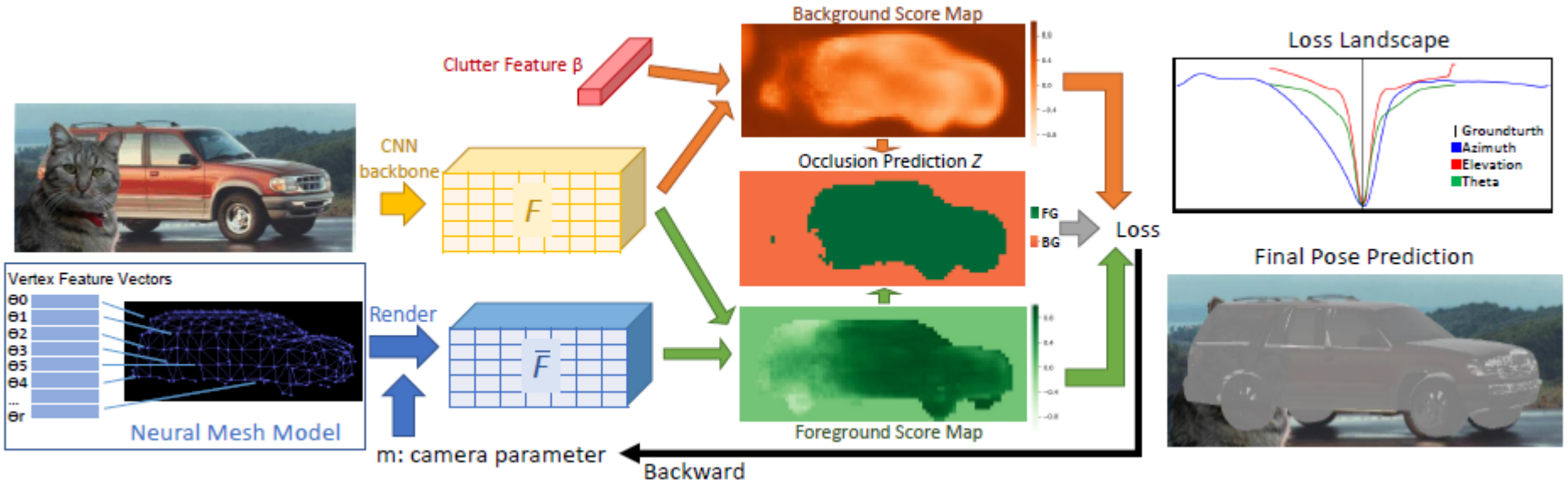
(4) Nemo: Traditional Render-and-Compare

- Basic Ideas: Generative on features gives invariance to unnecessary details of the objects.



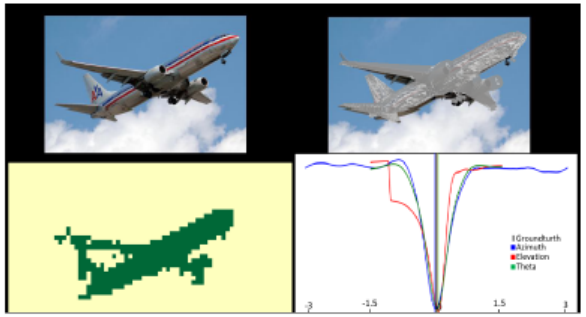
(4) NeMo: Generative Plus Outlier Process

- Outlier process makes NeMo robust to occluders. Generative on features smooths the energy landscape making search easier.

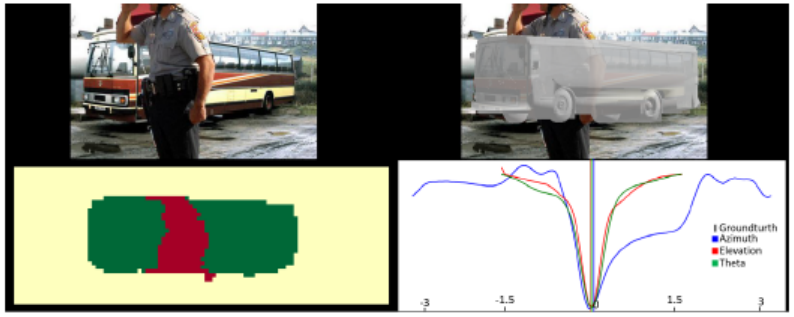


(4) NeMo: Results for 3D Pose estimation

- Also generalize to novel views.



(a) Aeroplane, L0



(b) Bus, L1

Evaluation Metric	$ACC_{\frac{\pi}{6}} \uparrow$				$ACC_{\frac{\pi}{18}} \uparrow$				$MedErr \downarrow$			
Occlusion Level	L0	L1	L2	L3	L0	L1	L2	L3	L0	L1	L2	L3
Res50-General	85.9	66.5	50.8	38.0	38.6	26.5	18.8	12.2	17.3	23.0	32.6	42.0
Res50-Specific	86.5	71.4	56.4	41.3	39.7	29.9	21.5	13.8	17.1	21.4	29.3	42.8
StarMap	**89.4**	71.1	47.2	22.9	59.5	34.4	13.9	3.8	9.0	17.6	34.1	63.0
NeMo	84.1	73.1	59.9	41.3	60.1	45.1	30.2	14.5	9.3	15.6	24.1	41.8
NeMo -MultiCuboid	86.7	**77.3**	**65.2**	**47.1**	**63.2**	**49.9**	**34.5**	**17.8**	**8.2**	**13.0**	**20.2**	**36.1**
NeMo -SingleCuboid	85.0	75.8	63.5	45.8	57.7	43.7	30.4	15.1	9.6	14.8	21.9	36.2

(4) NeMo: Summary

- This is a starting point for 3D CompNets. It needs to be refined by allowing more flexible geometric modeling.
- It also needs to have parts annotated.
- It needs to be applied to detection, classification, and part detection.
- Like all CompNets it needs to be extended to articulated objects, like humans and animals, whose non-rigidity makes them harder to model.